THE 1997 BBN BYBLOS SYSTEM APPLIED TO BROADCAST NEWS TRANSCRIPTION

Francis Kubala, Jason Davenport, Hubert Jin, Daben Liu, Tim Leek,
Spyros Matsoukas, David Miller, Long Nguyen, Fred Richardson,
Richard Schwartz, John Makhoul

BBN Systems and Technologies, Cambridge MA 02138

ABSTRACT

In this paper, we describe the BBN Byblos system used for the 1997 DARPA Hub-4 Broadcast News evaluation and discuss numerous improvements made to the system in 1997. We focused our effort entirely upon the two conditions containing studio-quality uncorrupted speech from native speakers, the so-called F0 (prepared speech) and F1 (spontaneous speech) conditions. In particular, we did not bother to create a separate acoustic model for narrow-band telephone speech. Our overall 1997 Hub-4 evaluation result was 20.4% WER, but our error rate on the F0/F1 conditions was only 14%. We ran regression tests on development test data that show we reduced word error rate by 22-30% on the F0/F1 conditions compared to our 1996 system. Sizable gains were achieved on all the other conditions as well, even though no extra effort was spent toward improving them. Brief summaries of three related efforts are also given covering the use of Byblos for Spanish news transcription, near real-time transcription, and automatic extraction of named entities from broadcast news.

1. INTRODUCTION

Broadcast news data is characterized by a bewildering variety of acoustic conditions since the speaker, speaking style, channel and environment change frequently. In 1997, we focused our attention entirely upon those portions of the news broadcasts containing studio-quality, uncorrupted speech from native speakers of American English. In the Hub-4 test context, data of this type is captured in two conditions labeled, F0 for prepared speech, and F1 for spontaneous speech. These conditions are typical of mainline news reportage over high-quality studio channels.

The motivation for this decision to concentrate our efforts is based in the observation that baseline speech recognition performance on clean wideband data is still unacceptably poor (approximately 25% WER in 1996) for most applications. Moreover, this kind of speech constitutes the majority of the content-rich portion of the news that would be most useful for information retrieval or extraction applications. We believe that fundamental improvements in recognition accuracy will be most easily achieved on uncorrupted data first, and that some fraction of any such fundamental improvements will translate to the corrupted conditions. Comparative tests described here seem to support this belief.

In the next section, we give an overview of the Byblos transcription system used for the 1997 Hub-4 evaluation. In section 2, we list the numerous improvements made to the system since the 1996 test. Results are discussed in section 3. In section 4, we briefly introduce three additional research efforts conducted with the 1997 Byblos transcription system. These include formal tests on Spanish broadcast news transcription, near real-time recognition of broadcast news, and an informal examination of named-entity extraction from automatic transcriptions of broadcast news.

2. SYSTEM DESCRIPTION

The system we used in the 1997 Hub-4 test differs only in the details from the 1996 system, which is described in [6]. Many details were changed, however, and these are discussed in the next section which describes the improvements we made in the past year. Below, we briefly review the salient general features of the Byblos transcription system that are common to our systems of the last 2 years.

The overall system organization for this year is identical to the 1996 Byblos transcription system:

1. Segment and classify gender
2. Cluster the segments
3. Decode with Speaker-Independent (SI) models, to get transcriptions for adaptation
4. Adapt models to each cluster
5. Decode with Speaker-Adaptively Trained (SAT) models, to produce the final answer

We use an efficient 2-pass decoder [11] to produce a set of N-best hypotheses from which we select the optimal answer by rescoring the N-best list with a more detailed model. The 2-pass decoder is run twice on the input data as noted above in steps (3) and (5). The first run produces the best SI hypothesis which is used as the supervisory transcription for unsupervised adaptation to the clustered test data. The second decode uses the adapted acoustic models to produce the final answers.

The monolithic broadcast news input is segmented and gender-classified in one step with a context-independent 2-gender phoneme decoder as described in [6]. The chopped segments are clustered automatically in an attempt to pool the data from each speaker for the benefit of unsupervised adaptation as described in [4]. The spectrum mean and variance is normalized over each segment, with speech and non-speech frames normalized separately. Gender-dependent acoustic models are estimated from the training data without regard to the speech environment or signal bandwidth

[12]. The gender-dependent SI models are refined by Speaker-Adapted Training (SAT) [1], [10], which attempts to model speaker differences jointly with the estimation of the speaker-independent phonetic model parameters.

For each gender, we create three models. The Phoneme-Tied Mixture (PTM) model has 45 phonetic codebooks with 256 Gaussians per phone and approximately 25K mixture weight vectors associated with the codebooks. The PTM model is used for the fast-match initial pass of the decoder and contains only within-word triphones.

The second pass of the Byblos decoder uses within-word State Clustered Tied Mixtures (SCTM) to generate a N-best list of the top scoring hypotheses. Each within-word SCTM codebook contains 64 Gaussians. There were 2K codebooks and 24K weight vectors created for the female model. The male model, which was estimated from twice as much data as the female model, had 4K codebooks and 38K weights vectors.

The top 300 best hypotheses from the second pass of the decoder are then rescored with a between-word SCTM model to select the top choice. Each between-word SCTM codebook has 64 Gaussians. The number of codebooks are approximately the same as for the within-word SCTM model, while the mixture weights increased in number by about 10%.

The language model was unchanged from 1996. We used about 450 million words of text from LDC corpora to estimate the trigram language model. Approximately one third of this data originated from broadcast news sources with the rest coming from newspaper sources. The acoustic training transcriptions were included 10 times in the LM training. The final LM had approximately 12M bigrams and 24M trigrams. Our recognition lexicon consisted of 45K words. Coverage of this lexicon on the 1996 Hub-4 test was 99.1%.

3. RECENT IMPROVEMENTS

Although the overall system didn't change much in the previous year, many small additive improvements were made to many parts of the system. The most important of these are described below.

3.1. Training Data

A new release of acoustic training data was published by LDC in 1997. This doubled the available training data compared to the 1996 release. The measured amount of usable speech in the complete 1997 training corpus is 80 hours. We observe an overall reduction in WER of 10% relative for this doubling of the training data. This translated to an absolute gain of about 3%. On closer examination, we see that most of the gain is due to the female speakers which constitute about one third of broadcast news data. We conclude that doubling the female training data from 13 to 26 hours has a significant effect, whereas increasing the male data from 26 to 52 hours is unimportant. We can conclude that 25 hours of training data from a given gender is an adequately sized training corpus.

We made an effort to segment the training data with greater care by cutting first at the longest pauses and the recursively cutting at the next longest while trying to minimize the variance in duration of the resulting segments. We also paid attention to the segment end conditions, ensuring that at least 10 frames of silence (background) were present. Although summary statistics for these effects greatly improved with this treatment of the training data, it had no effect on recognition performance.

3.2. Analysis

We improved our SNR-dependent cepstral normalization procedure by making better estimates of speech and noise frames and by including variance normalization as well. This accounted for an absolute gain of 0.7% in development tests.

We began using LPC smoothing of the Mel-warped spectrum this year and we observed that varying the numbers of LPC poles in the analysis had an inordinately large effect on WER – nearly 20% relative going from 14 poles to 28. Investigating further, we discovered that the 14 pole condition was degenerate and that an optimum LPC order occurred at 36 poles. We also discovered that LPC smoothing is sensitive to the spectrum floor which we usually pad to avoid very large negative excursions in the log-spectral domain. The gain for LPC smoothing was realized only after we reduced the spectum floor padding to 10^{-6}. On the male test speakers, the gain in WER was small but consistent – 0.7% absolute on the F0/F1 conditions, and about the same gain over all conditions. We ran out of time to test LPC smoothing on the female speakers and therefore use it only on the males in the 1997 Hub-4 evaluation.

3.3. Segmentation

Since the 1997 evaluation test came in the form of a single 3-hour waveform, we improvised a preliminary treatment to reduce the input into manageable chunk sizes. We chopped it arbitrarily into 8 arbitrary 23-minute pseudo episodes. As before, we segmented and gender-classified the input in one step with a context-independent 2-gender phoneme decoder as described in [6] But this year, we chopped more aggressively, creating segments that were only 4 seconds long on average. The shorter segments produced a better set of N-best hypotheses by reducing the number of permutations of the word errors, thereby extending the range between the best and worst N-best choices within the fixed length list. This alone gave us an absolute 0.5% gain.

We wanted to have some measure of segmentation quality independent of recognition results so we subjectively determined likely linguistic boundaries (sentence ends) and marked them in the transcription. We could then measure how often our automatic segmentation algorithms located boundaries at linguistically reasonable points. We found that our baseline segmentation performance, on average 8 second segments, was surprisingly good – 56% of the segment beginnings and 62% of their ends occurred at sentence boundaries. This segmentation accuracy was reduced somewhat for the shorter segments, but they improved performance nonetheless.

We found that there were many more filled pauses (err, ah, um, etc.) located at the beginning of segments and most of these were involved in errors. We modified the language model to relax the assumption that the segment begins at a linguistic sentence boundary, but we observed no additional gain in recognition performance.

3.4. Phone Model Estimation

Quinphone modeling has been shown to give small but consistent gains in large vocabulary speech recognition. Quinphone models are extended 2 contexts both to the left and the right of the phone to capture the coarticulation effect in more detail than the triphone. Since this extension increases the number of models significantly, a binary decision tree clustering is typically used to group similar model-states together to insure that there are sufficient data to train them.

For simplicity, we chose to capture only the boundary phone of the neighboring word in quinphone models that span word boundaries. We assign this boundary phone as the +2 or -2 context in the quinphone. A special word-boundary symbol plays the role of the +1 or -1 context. This simplifies the necessary change in the decoding modules of the quinphone system (i.e. there's not much difference compared to a triphone system). This simple implementation of the quinphone models gave us a solid 1 point gain in absolute WER reduction over the triphone system on the clean wideband speech conditions (F0 and F1 data) in the 1996 Hub-4 development test set.

We also changed our state clustering procedure to bootstrap from single Gaussians. This gave us another 0.7% absolute gain.

3.5. Pronunciation Modeling

We looked in depth at the behavior of our recognizer on spontaneous speech and found several effective ways to dramatically improve performance. Explicit modeling of filled pauses and laughter and extremely coarticulated phrases yielded a large gain of nearly 3% points. Details of this work are given in a companion paper in this volume, [8].

3.6. Adaptation

We improved our speaker adaptation procedures in several ways this past year. We changed our approach to Speaker Adapted Training (SAT) to a more computationally efficient Inverse Transform SAT that was introduced in [10]. We have also begun to use many diagonal transformations instead of a few full matrix transformations in SAT training. In the unsupervised adaptation stage, we are now using an iterative approach that starts with a constrained transformation which is relaxed on subsequent iterations. This work is described in detail in another paper in this volume, [5].

We also evaluated our speaker clustering algorithm and have improved it by introducing an additional penalty for clusters that contain only one short segment. Adapting to such clusters is unwise due to data sparsity. The new algorithm produces fewer singleton clusters and helps improve the quality of adaptation by a small margin.

3.7. Deleted Phone Modeling

During constrained decoding to create training labels, we often observed sequences of phones at the minimum duration of our HMM. We hypothesized that these could be signs of heavily coarticulated phonemes or completely deleted phones. We made several attempts to model phoneme deletions explicitly in the model but none of these yielded any gain. The one approach that had an effect was also the simplest; we added a skip transition from the first state to the last in our 5-state HMM. This reduced the minimum duration of the model to 2 frames, which apparently was enough to alleviate the problem. The number of phones at the minimum duration was reduced from about 12% to less than 3% for spontaneous speech and an absolute gain of 0.7% resulted. On further examination, we found that our PTM model produces many more phones at the minimum duration but the SCTM model does not have this problem. We verified that the fast-match pass using the PTM model does not produce search errors despite this suspicious behavior.

3.8. Vocal Tract Length Normalization

Vocal Tract Length Normalization (VTLN) has been repeatedly shown to be effective in reducing WER of the Switchboard/Callhome corpus, generally by more than 10% relative. Paradoxically, we were not able to reproduce this robust effect on the broadcast news corpus. We studied the problem extensively, but were not able to find an explanation for this difference in performance on the two domains.

In every detailed step of the procedure, VTLN applied to Broadcast News data appears to be functioning reasonably. The Gaussian Mixture Model (GMM) used to select the best stretch appears to work reasonably well for a wide range of data from a speaker (several minutes to as little as a few seconds). The distribution of stretches for the training speakers is approximately Gaussian, with a mean nearly centered at unity (no stretch). It also seems to work as well for clean data as for data corrupted with noise or music. The likelihoods produced by the GMM are correlated with the scores produced by the HMM during decoding. Nonetheless, the GMM is not selecting stretches that are correlated with word error rate.

We produced an oracle result by decoding each speaker at each quantized stretch and then determined the stretch required to achieve the lowest error rate by examination of the results. On Switchboard, the oracle result was relatively about 15% better than the no-VTLN result. For Broadcast News, the maximum gain available was only about 8% relative. More perplexing, however, was that all the potential oracle gain was lost in the fair test.

In the end, we were left with many conflicting results. We were not able to demonstrate any gain from VTLN on broadcast news data and did not use it in the 1997 evaluation.

3.9. Bigram on Mixture Weights

The usual Gaussian mixture model HMM allows a choice among the mixture components in a state for each frame. However, assuming that a particular token of a phoneme was all produced under some particular set of conditions (speaker, channel, environment, style), successive frames are highly correlated. Within our current HMM framework, the speech trajectory is captured by using derivatives of the input cepstra and the transition probability across states of the HMMs. In addition to these, we tried a novel approach to model the speech trajectory in the model space itself. We estimated the mixture weight probabilities conditioned on the mixture components from the previous frames. One can make an analogy to language modeling. The normal mixture weights represent a unigram model on the mixture sequence. We proposed to use a bigram model of the mixture sequence. This would directly model the dependence of each frame on the previous frame.

Using the bigram mixture model, we observed significant perplexity reductions for the bigrams versus the unigrams. For example, for a 32-compenent mixture density for the male gender, the perplexity is reduced almost 1/3 on the training data and 1/2 on the testing reference data. Moreover, the bigram perplexity increase for test over training is not very large — less than a factor of two. Despite this promising beginning, we were not able to achieve a gain of any significance on development tests.

3.10. Topic-Cache Language Model

We have developed a HMM topic indexer that is quite good at predicting reasonable topics for transcripts of general news stories. This is a domain in which the number of topics is very large (5000 or more) and very detailed. We attempted to use this capability to improve recognition results.

The general approach is to construct a cache of N-grams that have relevance to the story under hypothesis and raise their likelihoods in the general background language model. Last year we populated the cache with all words contained in the training stories labeled with the same topic(s). This was not successful, i.e. it failed to reduce word error rate, because it increased the likelihood of too many words, including many that were not relevant to the topic. We decided this year to try a more focused approach, increasing the weights only for those words whose probability is high given the topic. On average, this only affects a few hundred words per story.

But once again, were not able to achieve any significant gains for the adaptive LM cache model. The approach was limited by the relatively few potential corrections that could be made given the small number of topic keywords that were primed in the cache. Furthermore, most of these keywords needed to be correct for the topic to be triggered in the first place. In conclusion, we do not see a way to use topic information to increase recognition performance significantly.

4. DISCUSSION OF RESULTS

The Byblos system achieved a word error rate (WER) of 20.4% on the 1997 Hub-4 evaluation. In table 1, we display the WER for each condition on the output of the two decoding stages in our system. The first column contains the unadapted speaker-independent first-pass output. The overall SI performance of 22.6% is much better than we achieved on the 1996 development test set with any system, indicating that the 1997 evaluation test is substantially easier than the development test set or the 1996 evaluation set.

	Decoder Stage		relative
Condition	SI WER	SAT WER	gain
F0. prepared	13.2	12.3	7
F1. spontaneous	20.2	17.8	12
F2. low fidelity	38.5	32.6	15
F3. music	29.0	27.9	4
F4. noise	26.4	24.7	6
F5. non-native	28.9	28.2	2
FX. mixed	46.3	42.8	8
F0/F1	15.4	14.0	9
OVERALL	22.6	20.4	10

Table 1: 1997 Hub-4 core test results, showing relative gain between the SI and SAT-adapted recognition stages of the BYBLOS system.

The second column shows the final output of Byblos after unsupervised adaptation to the test data. Although every condition improves, the relative overall gain is the smallest we've observed since we began using unsupervised MLLR adaptation [9]. Moreover, the gains for adaptation are smaller on this test than we observed on either of the 1996 test sets. At this point, we have no satisfactory explanation for this observation.

	Byblos System		relative
Condition	1996 WER	1997 WER	gain
F0. prepared	22.8	18.9	17
F1. spontaneous	31.6	23.7	25
F2. low fidelity	34.3	30.7	11
F3. music	27.1	25.1	7
F4. noise	38.8	36.6	6
F5. non-native	38.1	35.8	6
FX. mixed	50.8	48.2	5
F0/F1	27.4	21.4	22
OVERALL	31.8	27.1	15

Table 2: Regression test on the 1996 Hub-4 UE data, showing relative gains achieved by the 1997 Byblos transcription system.

Table 2 shows results from a regression test performed on the 1996 Hub-4 evaluation data set. The results from the 1996 system shown are the official evaluation results published by NIST in 1996. The 1997 system results were generated on the same test set using the identical system and system parameters that were used in the 1997 Hub-4 evaluation. The relative gains achieved by the 1997 system shown in the last column clearly show the impact of our decision to concentrate all of our effort on improving core speech recognition represented by the clean wideband F0/F1 speech conditions. The largest gain by far was achieved for the spontaneous speech condition, which is very satisfying. For both the F0 and F1 conditions combined, our improvement for the year was 22%. On development test data, the F0/F1 improvement was 30% relative, from 27% WER down to 18%. Since these two conditions account for 70% of the data and since this portion of the data contains the most information-rich speech in broadcast news, this is a significant and useful achievement. Furthermore, since the gains for adaptation were greatly reduced on this test set, we consider the gains demonstrated here to be fundamental. That is, the basic model has been made more accurate and more robust at the same time.

After F0 and F1, the next largest gain was achieved for the low fidelity condition, F2, which includes narrow band telephone data. F2 data accounts for 16% of the 1997 test set. It is noteworthy that we achieved this gain entirely from general improvements in recognition developed on clean wideband data. As noted earlier, we declined to make a separate narrowband acoustic model or to study the F2 data explicitly in any way, so that we could channel our effort toward fundamental improvements. This result seems to affirm the efficacy of that approach. It's important to acknowledge that sizable additional gains can be made by explicitly modeling narrowband data as was done by most participants in the 1997 Hub-4 evaluation. We simply felt that the additional system complexity and research effort was not worth the expected return.

The relative gains were considerably smaller for conditions F3-FX. These 4 conditions account for only 19% of the test data. Still, these gains came for free and, excepting for the catch-all FX condition, performance loss for the degraded conditions is already less than 2 times the WER of clean prepared speech (F0). Concentrated research work on the degraded conditions seems counter-productive to us since our objective is to maximize the overall utility of automatic transcription of broadcast news for other applications.

4.1. Computational Resources

The computation for this evaluation was done on Intel-based PCs with 200MHz Pentium-Pro CPUs, 256MB of RAM, and 1.2GB of swap space. The operating system was Linux 2.0.3x and the compiler was GNU gcc 2.7.2 from the Free Software Foundation. Both of these are available as shareware at no cost. These machines have a 1995 SPEC base rating of 8.1 for the integer test, and 6.7 on the floating point test. The SPEC base tests restrict the allowable compiler optimizations to a standard set. For comparison, a Sun Sparc10 has SPEC95 base ratings of 1.0 for both the integer and floating point tests. This is the first year that we have attempted to run an evaluation on commodity PC hardware. In previous years, we were restricted to RISC architecture workstations running proprietary operating systems because the common PC was not up to the task of large vocabulary speech recognition. That era is now gone forever.

5. RELATED WORK

In addition to the Hub-4 Broadcast News evaluation in English, we have conducted additional work in three related areas involving broadcast news speech data.

5.1. Transcription of Spanish News

The Byblos recognition system used in the Hub-4NE Spanish evaluation is a simplified version of the one used in the Hub-4 English test. The Spanish system differed from the English system in the following ways:

- 2-level cepstal mean and variance normalization was not used
- quinphones were not used
- SAT models were not used
- fewer acoustic model parameters were used
- gender independent models were used

The dimensions of the reduced acoustic model are as follows. For the PTM model, we used 35 codebooks with 256 Gaussians each. For the SCTM model, we created only 1600 codebooks with 32 Gaussians each.

We used 27.5 hours of acoustic training data for the Spanish evaluation which came from the 81 episodes of development data provided by LDC. We held out the 3 episodes designated by NIST as the development test set. The data was processed in a manner similar to that used in the English system. The language modeling data (157 million words total) includes all of the LM data provided by LDC and also includes the transcriptions of the acoustic training data. The LM data were from newspaper sources and therefore required processing to transform numbers into words and to regularize acronyms, initials, and other orthographic forms typical of written language. We used a phonetic lexicon of Spanish from LDC and extended it via automatic transduction from letters to phonemes.

The results of the BBN Hub-4NE Spanish system in this year's evaluation indicate that an HMM system such as Byblos ports easily and effectively to other languages. Other than the work required to construct the phonetic dictionary, no language-specific knowledge, processing, or modeling were required to configure Byblos to handle Spanish input.

Test	Vocab	Adapt	WER	relative gain
Development	30K	no	28.1	
	40K	no	26.6	5%
	40K	yes	22.5	15%
Evaluation	40K	yes	19.9	

Table 3: Development and evaluation results for Byblos on the Hub-4 Spanish task.

In table 3, we show the effect of increased vocabulary and MLLR adaptation on WER for the development test set. The enlarged lexicon improved coverage on the held-out development test data. The 15% relative improvement for unsupervised adaptation to the test data is significantly larger that we ever observe on English. Part of the explanation for this effect may be due to the gender-independent acoustic model used here.

Remarkably, the absolute performance on Spanish is better than we achieved on the Hub-4 English test. Although we can't calibrate the difficulty of the different Hub-4 tests, it's clear that the Spanish result is a very good initial benchmark. There are several major limitations to the Spanish system, each of which should render it less accurate than the English system. The Spanish system had one third the training data for both the acoustic and language model compared to the English system. Moreover, the Spanish LM data came from newspapers rather than from news broadcasts. The Spanish acoustic model used far fewer parameters and the SAT paradigm was not used. The acoustic model was gender-independent. Furthermore, none of the recent improvements incorporated into the English system were used here and virtually no system development was done with the Spanish data other than verifying that the output appeared reasonable. All in all, this demonstrates that porting to languages other than English is straightforward and it does not require a language-specific research effort in order to succeed.

5.2. Toward Real-Time Transcription

We also submitted a formal contrast evaluation result to NIST for a system configured to run in six times the real-time duration of the speech being decoded. The resulting WER degraded only by 25% relative compared to the core evaluation system result which ran at an average rate of two hundred times real-time. These results are discussed in detail in [3], elsewhere in this volume.

5.3. Name Extraction from Speech

We have recently conducted a study of the effect of speech recognition errors and automatic transcription orthography (SNOR format) on the performance of our learned named-entity extraction engine, Nymble. Comparing to a baseline name extraction F-measure of 86.8 at 0% WER (ideal speech recognition), we observe 15% degradation in F-measure (to 73.4) when 20% WER transcriptions are used. We have found that name extraction performance is quite sensitive to WER and as such, this application is well suited for evaluating automatic transcription utility within an application framework. Details of this study are given in a companion paper, [7], elsewhere in this volume. A description of the Nymble system is described in [2].

Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency and monitored by NRaD under contract No. N66001-97-D-8501 and by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. Anastasakos, T., J. McDonough, R. Schwartz, "A Compact Model for Speaker-Adaptive Training", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
2. Bikel, D., S. Miller, R. Schwartz, R. Weischedel, "NYMBLE: A High-Performance Learning Name Finder", *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 194-201.
3. Davenport, J., L. Nguyen, R. Schwartz, F. Kubala, H. Jin, S. Matsoukas, D. Liu "Real Time Contrast System Description", *DARPA 1998 Broadcast News Transcription and Understanding Workshop*, Leesburg VA, Feb. 1998, elsewhere this volume.
4. Jin, H., F. Kubala, R. Schwartz, "Automatic Speaker Clustering", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.
5. Jin, H., S. Matsoukas, R. Schwartz, F. Kubala, "Fast Robust Inverse Transform SAT and Multi-stage Adaptation", *DARPA 1998 Broadcast News Transcription and Understanding Workshop*, Leesburg VA, Feb. 1998, elsewhere this volume.
6. Kubala, F., H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, J. Makhoul, "The 1996 BBN Byblos Hub-4 Transcription System", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.
7. Kubala, F., R. Schwartz, R. Stone, R. Weischedel, "Named Entity Extraction from Speech", *DARPA 1998 Broadcast News Transcription and Understanding Workshop*, Leesburg VA, Feb. 1998, elsewhere this volume.
8. Liu, D., L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, "Improvements in Spontaneous Speech Recognition", *DARPA 1998 Broadcast News Transcription and Understanding Workshop*, Leesburg VA, Feb. 1998, elsewhere this volume.
9. Leggetter, C. J., P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 110-115.
10. Matsoukas, S., R. Schwartz, H. Jin, L. Nguyen, "Practical Implementations of Speaker-Adaptive Training", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.
11. Nguyen, L., R. Schwartz, "Efficient 2-Pass Nbest Decoder", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.
12. Schwartz, R., H. Jin, F. Kubala, S. Matsoukas, "Modeling the F-Conditions (or not)", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997.